面向军事文本的命名实体识别

冯蕴天 张宏军 郝文宁
(解放军理工大学指挥信息系统学院 南京 210007)

摘 要 针对军事文本中的命名实体,提出一种基于条件随机场模型的半监督命名实体识别方法,旨在将人员军职军衔名、军事装备名、军用物资名、军事设施名、军事机构名(含部队番号)以及军用地名等军事命名实体的识别融合到一个统一的技术框架中。该方法针对军事文本的语法特点建立高效的特征集合,建立条件随机场模型对军事命名实体进行识别,并依次使用基于词典的方法和基于规则的方法对识别结果进行校正。实验表明,该方法在军事文本中能够出色地完成命名实体识别任务,在测试语料上的 F-值最高达到 90.9%,接近通用领域中命名实体识别的水平。

关键词 军事文本,命名实体识别,条件随机场,半监督学习,军事信息处理
中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.7.004

Named Entity Recognition for Military Text

FENG Yun-tian ZHANG Hong-jun HAO Wen-ning
(Institute of Command Information System, PLA University of Science and Technology, Nanjing 210007, China)

Abstract This paper presented a semi-supervised named entity recognition method based on conditional random field model for named entities in the military text, which aims at merging military named entities such as military appointment and military rank, military equipment, military supplies, military facilities, military institutions (including code designation) and military place names into a unified technical framework. The method establishes an efficient feature set according to the grammatical features of the military text, builds conditional random field model to identify the military named entities, and develops the method based on dictionary and the method based on rules to improve the results in turn. Experiments show that the method is able to complete the named entity recognition task in the military text well, and make F-value about testing the language material up to 90.9%, which is close to the level of named entity recognition in the commons area.

Keywords Military text, Named entity recognition, Conditional random field, Semi-supervised learning, Military information processing

1 引言

近年来,随着信息技术的发展以及部队作战指挥模拟化、网络化、信息化水平的迅速提升,我军进入高度信息化时代,大量的信息系统、数字化装备、模拟器材已投入到我军各层次的军事行动中,大量信息以电子文本的形式存在及被使用,海量的军事文本得以产生并仍以指数级的速度增长。如何处理并利用这些海量军事信息成为一个亟待解决的问题。

命名实体(Named Entity,NE)[1]是文本中基本的信息单位,是文本中的固有名称、缩写及其他唯一标识,是正确理解文本的基础。命名实体识别(Named Entity Recognition,NER)[2]是信息提取、问答系统、句法分析、机器翻译、面向 Semantic Web 的元数据标注等应用领域的重要基础性工作,在自然语言处理技术走向实用化的过程中占有重要地位。因此,对命名实体识别的研究具有重要的理论和现实意义,但命名实体识别在军事文本中未能得到较好的解决。本文针对军事文本的特点进行研究,提出一种面向军事文本的命名实体识别方法,这是下一步抽取文本中隐含的语义关系及军事行为知识的前提。

2 概述

2.1 军事文本

军事文本指军队在作战、训练及其他行动和工作中产生的以电子文本形式存在和使用的文档,是与军事相关的各种文本的统称。军事文本具体可分为军用文书和其他军事文本两类。军用文书[3]主要包括命令、通令、指示、通知、通报、请示、报告、计划、批复、公函、通告、布告等,其有统一的格式和书写方法,要求内容准确、简明扼要、严格保密;其他军事文本主要包括军事新闻文本、军事博客文本、军事评论文本等来源于互联网的文本,其他军事文本中存在口语化、网络化等现象,其句式不够严谨,语义表达不够清晰,又称为网络军事文本。军事文本中普遍存在大量的军事命名实体,它们有独特的语法结构,且构成都较为复杂,本文针对军用文书和网络军事文本进行命名实体识别研究,并提出一种统一的技术框架。

到稿日期:2014-07-19 返修日期:2014-10-24
冯蕴天(1990-),男,硕士生,主要研究方向为军用知识与数据工程,E-mail:fengyuntian2009@live.cn;张宏军(1963-),男,教授,博士生导师,主要研究方向为军事建模与仿真;郝文宁(1971-),男,博士,副教授,主要研究方向为军用知识与数据工程。

2.2 军事命名实体

命名实体识别在通用领域的研究已经比较成熟，如在新闻领域命名实体识别系统的 F-值（系统性能的评估参数）可达到 90%以上[4]，接近人类识别水平。在通用领域中，命名实体一般具有数目相对稳定、结构比较规范、命名规则比较统一等有利特点，主要包括[5]人名、地名、组织机构名等。但在刚进入信息化阶段的军事领域，命名实体识别的研究却不够深入。

军事命名实体是指在军事文本中与军事相关的各种命名实体的统称，主要包括人员军职军衔名、军事装备名、军用物资名、军事设施名、军事机构名（含部队番号）、军用地名等。它们的产生往往以更为系统、复杂的军事知识为依据，其构成模式也更为多样，对它们的识别必须兼顾其语言规律和军事特性。

3 面向军事文本的命名实体识别方法

3.1 预处理

对军事文本进行预处理，不仅能直接提取重要格式信息，还可以消除噪声，减少干扰，完成特征提取前文本的准备工作，提高命名实体识别的效率。对军事文本的预处理可分为以下 4 个步骤。

3.1.1 提取重要格式信息

军用文书有统一的格式和书写方法，有简明的标题、编号、秘密等级、主题词、主送和抄送单位、制发机关、承办人和日期、时间、署名等格式信息。从军用文书中提取的格式信息可直接成为重要的军事命名实体，或包含重要的军事命名实体。格式信息中的命名实体是军用文书的中心词，对文书的自动理解具有重要意义。

3.1.2 文本规范化

网络军事文本中包含许多不规范的词或符号，这些词或符号对命名实体的识别没有实际意义，对文本进行规范化处理，既不会对文本整体文义产生影响，又能减少大量干扰信息。文本规范化主要包括以下几种方式：进行繁简字转换，如“中華”转换为“中华”；清除连续的符号组合，如“＝＿＝”、“（ ╰＿╯ ）＃”、“↖(^ω^)↗”等，这些符号组合在军事博客文本中代表某种表情；去除干扰词，如“赞”、“踩”、“OK”、“Thanks”、“Good”等，这些词在军事评论文本中并不包含实际意义。

3.1.3 文本分词

本文采用中科院研制的汉语词法分析系统 ICTCLAS 对军事文本进行分词处理。ICTCLAS 是国内最权威的中文分词系统，在《人民日报》语料上学习的分词器，可同时实现中文分词和词性标注，词性标注的结果可直接作为下一步统计学习的重要特征。ICTCLAS 可以使军事文本保持适中的分词粒度，若分词粒度过小，最终特征集的规模则会急剧增加，使统计学习模型计算收敛过慢；若分词粒度过大，则不仅需要在分词时加入完备的军事领域词典，在学习过程中还易出现“过拟合”现象。

3.1.4 训练语料的转换

为了对统计学习模型进行训练，本文引入 13 种符号来表示每个输入单元，完成训练语料的转换。其中，PB 表示人员军职军衔名的开始，PI 表示人员军职军衔名的内部，WB 表示军事装备名的开始，WI 表示军事装备名的内部，MB 表示军用物资名的开始，MI 表示军用物资名的内部，FB 表示军事设施名的开始，FI 表示军事设施名的内部，AB 表示军事机构名（含部队番号）的开始，AI 表示军事机构名（含部队番号）的内部，LB 表示军用地名的开始，LI 表示军用地名的内部，O 表示其他。

3.2 基于条件随机场的命名实体识别

3.2.1 条件随机场

条件随机场（Conditional Random Field，CRF）是由 Lafferty[6]等人于 2001 年提出的一种判别式概率模型，其特点是假设输出随机变量构成马尔可夫随机场，常用于分词、命名实体识别等预测问题。

条件随机场定义：设 X 与 Y 是随机变量，$P(Y|X)$是在给定 X 的条件下 Y 的条件概率分布，若随机变量 Y 构成一个由无向图 $G=(V,E)$表示的马尔可夫随机场，即

$$P(Y_v|X,Y_u,u\neq v)=P(Y_v|X,Y_w,w\sim v)$$

对任意结点 v 成立，则称条件概率分布 $P(Y|X)$为条件随机场。其中 $u\neq v$ 表示结点 v 以外的所有结点 u；$w\sim v$ 表示在图 $G=(V,E)$中，与结点 v 有边连接的所有结点 w；Y_v、Y_u 与 Y_w 为结点 v、u 与 w 对应的随机变量。

条件随机场的参数化形式：设 $P(Y|X)$为条件随机场，则在随机变量 X 取值为 x 的条件下，随机变量 Y 取值为 y 的条件概率具有如下形式：

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

其中，

$$Z(x)=\sum_{y}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

其中，t_k 和 s_l 是特征函数，λ_k 和 μ_l 是对应的权值，$Z(x)$是规范化因子，求和是在所有可能的输出序列上进行的。

CRF＋＋是著名的条件随机场开源工具，也是目前综合性能最佳的 CRF 工具。本文使用 CRF＋＋－0.58 工具包实现条件随机场模型的定制，通过对连续数据进行标注进而实现命名实体的识别。

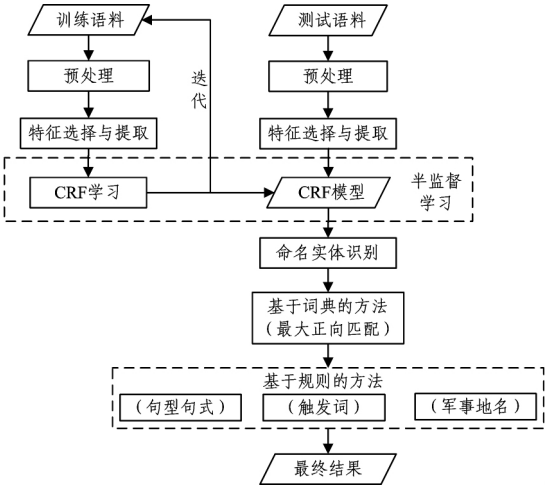

图 1 面向军事文本的命名实体识别技术框架

本文基于条件随机场模型提出的面向军事文本的命名实体识别技术框架，如图 1 所示。该框架可同时适用于军用文书和网络军事文本，并能针对多种不同军事命名实体进行识别。首先，对训练语料和测试语料进行预处理，并对预处理后的语料进行特征选择与提取，必须保证在训练语料和测试语

料中提取相同的特征并且顺序一致；然后在训练语料上进行条件随机场模型的学习，产生条件随机场模型，并使用半监督学习的方法对模型进行反复迭代，最终得到最优的条件随机场模型；最后利用最优的条件随机场模型在测试语料中进行命名实体识别，产生初步的识别结果，并依次使用基于词典的方法和基于规则的方法对初步的识别结果进行校正，得到最终的识别结果。

3.2.2 特征选择及特征模板的构建

在上文所述的条件随机场的参数化形式中，$t_k(y_{i-1},y_i,x,i)$是定义在边上的特征函数，称为转移特征，依赖于当前和前一个位置；$s_l(y_i,x,i)$是定义在结点上的特征函数，称为状态特征，依赖于当前位置。特征函数 t_k 和 s_l 取值为 1 或 0；当满足特征条件时取值为 1，否则为 0。条件随机场模型的优点在于它可以融合多种上下文特征，本文使用以下 5 种特征建立特征函数。

(1)词特征

文本分词后产生的每个词本身作为一种特征，词特征能够较完整地反映文本的基本信息。如果训练语料规模足够大，仅使用词特征就能较好地实现命名实体的识别。

(2)词性特征

在文本分词过程中同时对每个词进行词性的标注，大量研究表明[7]，把词性作为一种特征建立条件随机场模型可显著地提高命名实体识别的性能。

(3)英文字母、短横线及数字的组合特征

军事装备名通常包含英文字母、短横线及数字的组合，如“ZTZ-99 式主战坦克”等。因此把英文字母、短横线及数字的组合作为一种特征有助于军事装备名的识别。

(4)左、右边界词特性

通过对大量军事文本进行语法分析，发现军事命名实体的前面和后面出现一些特定动词的概率很高，这些动词称为命名实体的左、右边界词，因此必须建立完备的左、右边界词库。如“5 月下旬，某炮兵旅与某陆航部队联合进行的一场空地立体火力打击演练拉开战幕。”、“今年 3 月下旬，该市永定区牧塔村突发森林火灾，军分区迅速组织机关勤务队和永定区民兵应急分队赶赴火场进行扑救。”等，其中的“联合”、“组织”既可作为左边界词又可作为右边界词。

(5)中心词特性

复合军事命名实体中通常包含一些特定名词，这些词称为命名实体的中心词，中心词的出现很大概率上预示着军事命名实体的出现，因此必须建立完备的中心词库。如“中方海军航空兵某团 8 架‘飞豹’战斗机将主要扮演蓝军，已全部准备就绪，随时可投入战斗。”，其中的“航空兵”、“战斗机”为中心词。

特征模板是指特征的联合表示，可对多种上下文信息进行综合。本文将上述 5 种特征进行融合，建立了多种特征模板，对文本中词的特征进行全面表达。本文将特征模板的活动窗口大小设置为 3，即需要包含当前词的前两个词和后两个词的特征。若活动窗口过小，会使特征表达不全面，从而降低命名实体识别的性能；若活动窗口过大，则会产生庞大的特征空间，会使模型计算收敛过慢。仅使用单一特征建立的模板称为原子特征模板，如表 1 所列；结合多种上下文特征建立的模板称为复合特征模板，如表 2 所列。其中，W 代表词特征，P 代表词性特征，S 代表英文字母、短横线及数字的组合特征，B 代表左、右边界词特性，C 代表中心词特征，下标 0、−1、−2、1、2 分别代表当前词、当前词左边的第一个词、当前词左边的第二个词、当前词右边的第一个词、当前词右边的第二个词。

表 1 原子特征模板

序号	模板	特征模板含义
1	W_0	当前词
2	W_{-1}	当前词左边第一个词
3	W_{-2}	当前词左边第二个词
4	W_1	当前词右边第一个词
5	W_2	当前词右边第二个词
6	P_0	当前词的词性
7	P_{-1}	当前词左边第一个词的词性
8	P_{-2}	当前词左边第二个词的词性
9	P_1	当前词右边第一个词的词性
10	P_2	当前词右边第二个词的词性
11	S_0	当前词是否为英文字母、短横线及数字
12	S_{-1}	当前词左边第一个词是否为英文字母、短横线及数字
13	S_1	当前词右边第一个词是否为英文字母、短横线及数字
14	$B_{-1,1}$	当前词左、右边第一个词是否为左、右边界词
15	$B_{-2,2}$	当前词左、右边第二个词是否为左、右边界词
16	C_0	当前词是否为中心词

表 2 复合特征模板

序号	模板
1	$W_{-2},W_{-1},W_0,W_1,W_2$
2	$P_{-2},P_{-1},P_0,P_1,P_2$
3	$W_{-2},W_{-1},W_0,W_1,W_2,P_{-2},P_{-1},P_0,P_1,P_2$
4	$W_{-2},W_{-1},W_0,W_1,W_2,P_{-2},P_{-1},P_0,P_1,P_2,S_{-1},S_0,S_1$
5	$W_{-2},W_{-1},W_0,W_1,W_2,P_{-2},P_{-1},P_0,P_1,P_2,S_{-1},S_0,S_1,B_{-1,1},B_{-2,2}$
6	$W_{-2},W_{-1},W_0,W_1,W_2,P_{-2},P_{-1},P_0,P_1,P_2,S_{-1},S_0,S_1,B_{-1,1},B_{-2,2},C_0$

3.3 半监督学习 Self-Training 算法

半监督学习[8]是监督学习和无监督学习相结合的一种统计机器学习方法，可利用少量的标注语料和大量的未标注语料进行训练和分类。军用文书属于涉密文件，对军用文书的标注需要军方保密部门的参与和监管，从而使得标注语料获取的难度增大；网络军事文本不属于涉密文件且易于大量获取，但对海量存在的网络军事文本进行标注，同样会消耗大量的人力、物力资源。若仅使用少量标注语料进行有监督学习，通常难以学得泛化能力强的模型，故本文使用半监督学习 Self-Training 算法在少量标注语料和大量未标注语料上学得泛化能力较强的模型，图 2 为半监督学习 Self-Training 算法的流程图。

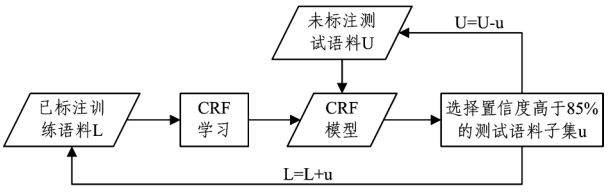

图 2 半监督学习 Self-Training 算法的流程

首先，用已获取的少量已标注训练语料 L 对条件随机场模型进行训练，产生条件随机场模型 C_0；然后，使用模型 C_0 对未标注测试语料 U 进行命名实体识别，并对测试语料 U 的标注结果进行置信度估算，上文中条件随机场 $P(Y|X)$被称为样本标注序列 Y 的后验概率，它反映了在看到数据样本 X 后 Y 成立的置信度，因此可用 $P(Y|X)$表示条件随机场模型

对测试语料标注结果的置信度，其计算方法同上文的条件随机场模型计算公式；最后，选择测试语料 U 中置信度高于85%的子集 u，并把 u 加入到训练语料 L 中，同时从测试语料 U 中删除 u；按照上述3个步骤进行迭代，依次产生条件随机场模型 C_1，C_2，C_3，…，当本次迭代出的条件随机场模型 C_k 对测试语料中所有文本标记的置信度与模型 C_{k-1} 对测试语料中所有文本标记的置信度的差异均小于0.5%时，迭代停止，可认为模型达到收敛状态，从而产生泛化能力最优的条件随机场模型 C_k。

3.4 识别结果校正

军事命名实体的构成存在特定规律，并且来源相对固定。因此，本文采用基于词典的方法和基于规则的方法对模型 C_k 的命名实体识别结果进行校正，进一步提高命名实体识别的性能。

3.4.1 基于词典的方法

提取《军语》中综合、国防、作战(综合)、指挥、陆军、海军、空军、第二炮兵等章节的词条建立军事命名实体基本库。《军语》中包含的军事术语是军队在作战、训练及其他行动和工作中统一使用的规范化的军事用语，是组织各项军事活动、规范军队体制编制、表示各种武器装备、设施名称及功能、撰写各类军用文书的标准用语，大量军事命名实体都来源于《军语》。

对目前已存在的军事训练本体库中的军事概念进行提取，建立军事训练命名实体基本库。军事训练本体库是我军军事训练领域本体建模的主要成果之一，可从中提取出大量规范的军事训练领域命名实体。

人工收集我军及外军军职军衔名称，建立人员军职军衔命名实体库，词条总计142个，如“师长”、“作训参谋”、“准将”等；人工收集我军及外军著名武器装备名称，建立军事装备命名实体库，词条总计764个，如“歼-10战斗机”、“东风-21中程弹道导弹”、“F-22‘猛禽’”等；人工收集我军及外军常用物资名称，建立军用物资命名实体库，词条总计87个，如“被装”、“粮秣”、“油料”等；人工收集我军及外军主要设施名称，建立军事设施命名实体库，词条总计73个，如“军用码头”、“军用铁路专线”、“试验场”等；人工收集我军及外军重要机构名称，建立军事机构命名实体库，词条总计115个，如“总参谋部”、“军事科学院”、“美国参谋长联席会议”等。

上述命名实体库共同构成了军事命名实体词典，词典中的每一个词条代表着一个命名实体，依据该词典可对条件随机场模型 C_k 标注的结果进行校正。使用词典在模型 C_k 标注的语料中进行最大正向匹配，语料由分词处理后的军事文本构成。假设词典中最长的词条字符数为 m，然后取军事文本的前 m 个字符，在词典里查找是否存在与之完全匹配的词条，若存在，说明该词是某类军事命名实体，故在文本中将该词去掉；若不存在，则去掉这 m 个字符的最后一个字符，检查是否为单字符，若是，则在文本中去掉此字符，继续取文本的前 m 个字符反复循环，若不是，则继续判断词典中是否存在与之完全匹配的词条，如此反复循环。当与某词条实现完全匹配时，判断这些字符是否已被模型 C_k 标注为同一个军事命名实体，若不是，则按照该词条在词典中所属的类别对这些字符进行标注；若是，则继续进行匹配。

3.4.2 基于规则的方法

通过人工总结出一些启发式的句法、词法和语法规则，制定出规则模板。使用规则模板对基于词典的方法产生的校正结果进行“二次”校正。

(1)句型句式规则

军用文书中经常使用大量固定的句型句式，主要包括介词结构[3]等，如“为了……特指示如下”、“由……特制定本办法”、“根据……特报告如下”、“据近查……特作如下指示”等。利用这些介词结构来充当句子的定语、状语、补语，以表示动作行为的目的、范围、依据、方向、对象、时间、地点等，从而使叙述更加周密、明确。军事命名实体经常嵌套在上述结构中，因此可通过对固定句型句式的识别来判定军事命名实体。

(2)触发词规则

军事命名实体中经常包含一类固定的词，这些词的出现通常预示着军事命名实体的出现，这些词称为触发词，如“XX团”、“XX导弹”、“军事XX”中的“团”、“导弹”、“军事”为触发词，它们不仅有着鲜明的军事特征，还标识着军事命名实体的类别。因此，本文建立了完备的触发词知识库来判定军事命名实体。

(3)军事地名规则

军事地名是军用文书中必不可少的命名实体，其出现形式通常包含着具体的坐标或者具体的方位。如“小曹庄(66，87)”、“炮台山北麓”，利用这些坐标和方位词充当定语，以对军事地名的表述更精确、完整，并且坐标和方位词可作为军事地名的一部分。因此可通过对坐标和方位词的识别来判定军事地名。

4 实验结果及分析

4.1 实验设置

由于目前没有比较权威统一的军事语料库，因此采用人工收集的方式构建军事文本库。实验样本包括100篇战斗文书、100篇执勤文书，共200篇军用文书，总计175435字；还包括100篇收集于《解放军报》的军事新闻、100篇收集于主流门户网站的军事新闻、100篇军事博客、100篇军事评论，共400篇网络军事文本，总计367835字。将实验样本中的军用文书和网络军事文本各取80%作为训练语料，其余20%作为测试语料。本文针对不同类型的军事文本设置3组对比实验。

实验1 直接在训练语料上进行条件随机场模型的学习，利用得到的条件随机场模型分别对测试语料中的军用文书和网络军事文本进行命名实体识别，并对识别结果进行评估，该实验方法可记为CRF。

实验2 使用半监督学习Self-Training算法在训练语料进行条件随机场模型的迭代学习，利用最优的条件随机场模型分别对测试语料中的军用文书和网络军事文本进行命名实体识别，并对识别结果进行评估，该实验方法可记为CRF+Self-Training。

实验3 使用半监督学习Self-Training算法在训练语料上进行条件随机场模型的迭代学习，利用最优的条件随机场模型分别对测试语料中的军用文书和网络军事文本进行命名实体识别，并依次使用基于词典的方法和基于规则的方法对初步的识别结果进行校正，得到最终的识别结果并进行评估，该实验方法可记为CRF+Self-Training+两种校正方法。

(下转第47页)

[6] Japaridze G. From truth to computability II [J]. Theoretical Computer Science,2007,379:20-52
[7] Japaridze G. In the beginning was game semantics. Games:Unifying Logic,Language and Philosophy[C]// Majer O,Pietarinen A-V,Tulenheimo T,eds. Springer,2009:249-350
[8] Japaridze G. The propositional logic of elementary tasks [J]. Notre Dame Journal of Formal Logic,2000,41(2):171-183
[9] Sipser M. Introduction to the Theory of Computation [M]. Thompson,2006
[10] Xu W,Liu S. Deduction theorem for symmetric cirquent calculus [J]. Advances in Intelligent and Soft Computing,2010,82:121-126
[11] Blass A. A game semantics for linear logic[J]. Annals of Pure and Applied Logic,1992,56:183-220
[12] 朱大铭,马绍汉. 算法分析与设计[M]. 北京:高等教育出版社,2009
Zhu Da-ming,Ma Shao-han. Design and Analysis of Algorithms [M]. Beijing:Higher education press,2009

(上接第 18 页)

4.2 实验结果

本实验使用 3 个指标来衡量命名实体识别的性能:正确率、召回率、F-值。其计算公式如下:

$$\text{正确率}(P)=\frac{\text{系统正确识别的实体个数}}{\text{系统识别的实体个数}}\times 100\%$$

$$\text{召回率}(R)=\frac{\text{系统正确识别的实体个数}}{\text{文档中的实体总数}}\times 100\%$$

$$F\text{-值}=\frac{2\times P\times R}{P+R}\times 100\%$$

对实验 1、实验 2、实验 3 的命名实体识别结果进行正确率、召回率、F-值的计算,结果如表 3 所列。

表 3 实验结果

实验	文本类型	实体总数	识别个数	正确个数	正确率/%	召回率/%	F-值/%
CRF	军用文书	663	644	559	86.80	84.31	85.54
	网络军事文本	1045	1019	869	85.28	83.16	84.21
CRF+Self-Training	军用文书	663	647	579	89.49	87.33	88.40
	网络军事文本	1045	1017	892	87.71	85.36	86.52
CRF+Self-Training+两种校正方法	军用文书	663	646	602	93.19	90.80	91.98
	网络军事文本	1045	1027	929	90.46	88.90	89.67

实验 1 仅使用了条件随机场模型,在军用文书中识别的正确率、召回率、F-值仅为 86.8%、84.31%、85.54%,在网络军事文本中识别的正确率、召回率、F-值仅为 85.28%、83.16%、84.21%,效果不太理想,没有达到进行军事应用的标准。

实验 2 使用了 Self-Training 算法对条件随机场模型进行迭代学习,在军用文书中识别的正确率、召回率、F-值分别为 89.49%、87.33%、88.4%,在网络军事文本中识别的正确率、召回率、F-值为 87.71%、85.36%、86.52%,该方法解决了大量标注语料难以获取的问题,提高了识别的性能。

实验 3 在实验 2 的基础上依次采用基于词典的方法和基于规则的方法对识别结果进行校正,在军用文书中最终识别的正确率、召回率、F-值可达到93.19%、90.8%、91.98%,在网络军事文本中最终识别的正确率、召回率、F-值可达到 90.46%、88.9%、89.67%,使得在军事文本中的命名实体识别性能达到与通用领域相当的水平。实验 3 的方法在继承了基于统计学习模型方法的基础上同时吸收了基于词典和基于规则方法的优点,其在军用文书中识别的正确率、召回率、F-值相比于实验 2 分别提高了 3.7%、3.47%、3.58%,在网络军事文本中识别的正确率、召回率、F-值相比于实验 2 分别提高了 2.75%、3.54%、3.15%,训练语料相对于海量的军事文本来说数量明显不足。采用两种校正方法可准确识别出更多的复合词和嵌套词,使识别的正确率和召回率都有大幅度的提高。

上述 3 组对比实验中,对军用文书识别的性能始终略高于对网络军事文本识别的性能,其主要原因在于军用文书文体较为规范,而网络军事文本中普遍存在较多口语化、网络化的词汇,因而为命名实体识别造成较大困难。

结束语 本文对军事文本和军事命名实体进行定义和研究,并针对其特点提出了一种面向军事文本的命名实体识别方法。该方法不仅同时使用多种原子特征模板和复合特征模板进行特征的表达,还采用了半监督学习 Self-Training 算法对条件随机场模型进行迭代学习。通过对比在人工搜集的军事文本上进行的 3 组实验结果,表明该方法能够有效解决军事文本中的命名实体识别问题,并能获得较好的识别性能。如何进一步提高命名实体识别的速度,是下一步研究的重点。

参 考 文 献

[1] Tjong Kim Sang E F,De Meulder F. Introduction to the CoNLL-2003 shared task:Language-independent named entity recognition[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics,2003:142-147
[2] McCallum A,Li W. Early results for named entity recognition with conditional random fields,feature induction and web-enhanced lexicons[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003-Volume 4. Association for Computational Linguistics,2003:188-191
[3] 向音. 军用文书的语篇特征初探[J]. 办公室业务,2011,10:19-20
Xiang Yin. The preliminary study about discourse features in the military document[J]. Office Operations,2011,10:19-20
[4] 邱泉清,苗夺谦,张志飞. 中文微博命名实体识别[J]. 计算机科学,2013,40(6):196-198
Qiu Quan-qing,Miao Duo-qian,Zhang Zhi-fei. Named entity recognition on chinese microblog[J]. Computer Science,2013,40(6):196-198
[5] 张晓艳,王挺,陈火旺. 命名实体识别研究[J]. 计算机科学,2005,32(4):44-48
Zhang Xiao-yan,Wang Ting,Chen Huo-wang. Research on named entity recognition[J]. Computer Science,2005,32(4):44-48
[6] 李航. 统计学习方法[M]. 北京:清华大学出版社,2012:191-210
Li Hang. statistical learning methods[M]. Beijing:Tsinghua University Press,2012:191-210
[7] Nadeau D,Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes,2007,30(1):3-26
[8] 黄鸿,李见为,冯海亮. 基于半监督流形学习的人脸识别方法[J]. 计算机科学,2009,35(12):220-223
Huang Hong,Li Jian-wei,Feng Hai-liang. Face recognition based on semi-supervised manifold learning[J]. Computer Science,2009,35(12):220-223